

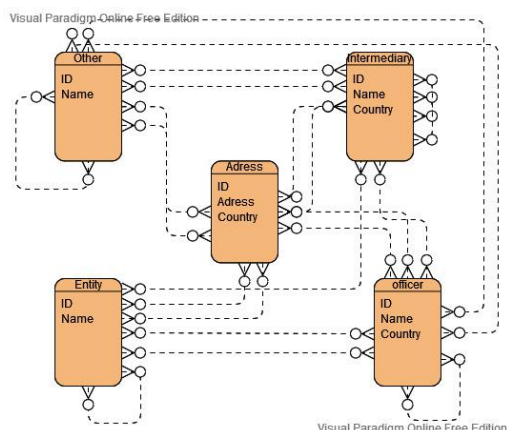
Aufgabenstellung MC2

Tobias und Yannic sind zwei junge Journalisten und arbeiten zusammen seit einigen Monaten bei der Nord-West Schweizer Zeitung. Eines Morgens checkt Tobias sein anonymes Postfach im Internet und stösst auf 6 csv-Dateien. Nach genauerer Betrachtung der einzelnen Dateien vermutet Tobias, dass es sich dabei um etwas Riesiges handeln könnte! Die Daten sind bis auf einige Kleinigkeiten bereits feinsäuberlich aufbereitet, darin enthalten sind in 5 Unterschiedlichen Tabellen Adressen, Branchen sowie Personen- und Firmennamen. Tobias holt rasch seinen Kollegen Yannic und die beiden beschliessen damit in die IT zu gehen. Andreas der IT-Fachmann für Datenbanken erkennt sofort um was es sich dabei handelt. Der mysteriöse Whistleblower hat die Daten feinsäuberlich für eine Graphen Datenbank aufbereitet. «Wieso er wohl ausgerechnet so eine komplizierte Datenbank verwendete?», Fragten sich Tobias und Yannic. «Das will ich euch mal genauer erklären», antwortete Andreas.

Nebst der Tatsache, dass in diesen Daten die ganzen Beziehungen schon feinsäuberlich aufbereitet sind, eignen sich Graph-Datenbanken hervorragend, um Operationen mit Beziehungen im Mittelpunkt durchzuführen.

Doch Andreas empfiehlt den beiden erst einmal die Daten der csv-Dateien als Konzeptionelles Modell zu modellieren, damit man einen besseren Überblick erhält.

Tobias erinnerte sich noch an den Aufbau eines Konzeptionellen Modelles und beginnt dieses zu Zeichnen:



Tobias ist hochmotiviert und möchte sofort alle Skandale aufdecken, welche der Datensatz beinhaltet. Doch er muss schnell feststellen, dass er nicht ganz genau weiss, wie er die Skandale aufdecken kann.

Andreas erläuterte, dass Graph Datenbanken nicht sofort die Lösung bietet, sondern man einen Startpunkt wählen muss.

Yannic erinnert sich daran, dass die FIFA ein guter Anhaltspunkt sein könnte, da diese immer wieder in

diverse Skandale verwickelt sind.

Tobias überprüft mittels Regex, ob sich die Fifa in dem Datensatz befindet und wird fündig. Die Firma heisst «FIFA INTERNATIONAL LIMITED».

So entschliessen sich Yannic und Tobias, dass sie herausfinden möchten, mit was für Instanzen die Fifa in Verbindung steht.

Andreas meint, dass das eine exzellente Idee sei, da eine Graph Datenbank die einzige Datenbank ist, welche eine solche Frage effizient beantworten kann.

Somit beginnen Yannic und Tobias die csv-Dateien mittels «Neo4j admin importer» in eine Neo4j Datenbank einzulesen:

```
neo4j-admin import --database=leaks --delimiter="," --skip-duplicate-nodes --
nodes=Address=import/nodes-addresses.csv --nodes=Entity=import/nodes-
entities.csv --nodes=Intermediary=import/nodes-intermediaries.csv --
nodes=Officer=import/nodes-officers.csv --nodes=Other=import/nodes-others.csv
--relationships=import/relationships.csv
```

Die Datenbank war in wenigen Sekunden aufgebaut, was selbst Andreas verwunderte.

Um nachzuprüfen, ob beim Konzeptionellen Modell nichts vergessen wurde, haben die beiden noch schnell das logische Modell der Datenbank mit dem Befehl

```
call db.schema.visualization()
```

generieren lassen:

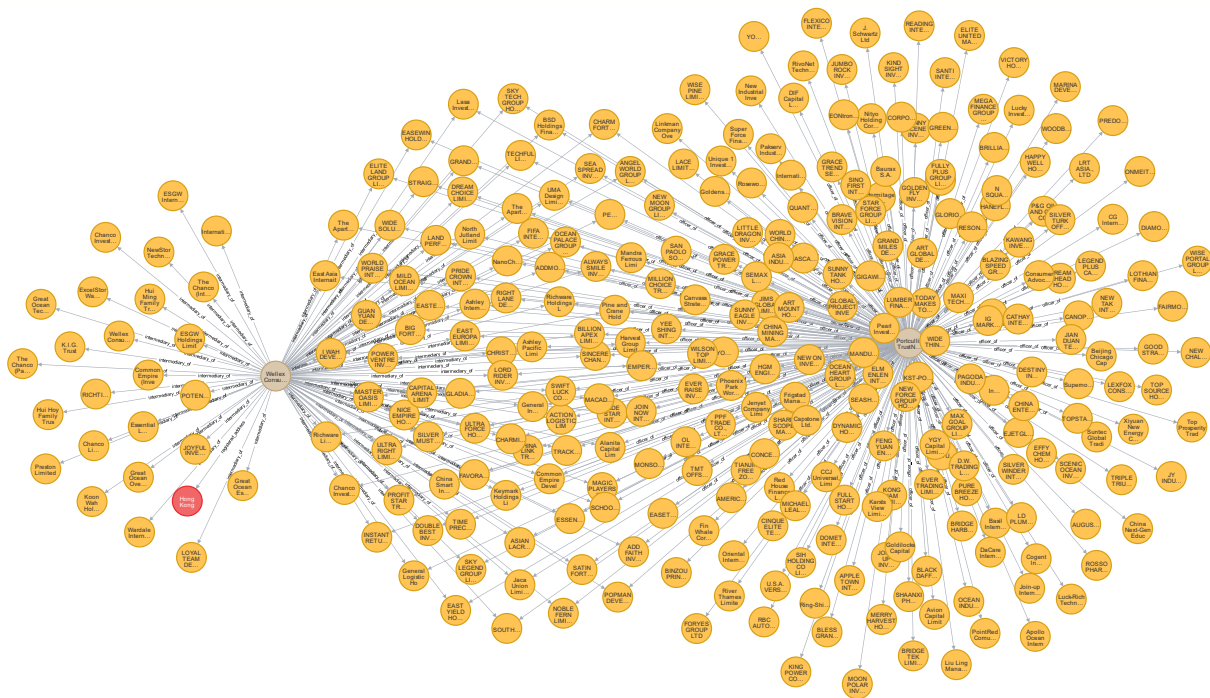


Nachdem sich Tobias versichert hatte, dass das Konzeptionelle Modell stimmt, führte Yannic die Abfrage aus, welche die Geheimnisse der Fifa enthüllen sollten:

```
match(a:Entity { name: "FIFA INTERNATIONAL LIMITED" })-[*..2]-(b) return *
```

Dieser Befehl zeigt alle Nodes und Kanten, ausgehend von der Fifa mit der maximalen Relationstiefe 2.

Das Resultat sieht so aus:



Yannic und Tobias erkennen sofort, dass die Vermittler «Wellex Consultancy Limited» und «Portcullis TrustNet (BVI) Limited» eine Entscheidende Rolle spielen könnten (graue Nodes in Darstellung).

Tobias bemerkt, dass die kürzeste Verbindung zwischen den beiden Vermittlern auch von Bedeutung sein könnte.

Da mit einer solchen Datenbank eine Abfrage dieses Types auch sehr effizient ausgeführt werden kann, dauert es nicht lange, bis ein Resultat vorliegt:

Abfrage:

```
match(a:Entity { name: "FIFA INTERNATIONAL LIMITED" })-[*..2]-(b) return *
```

Resultat:



Somit ist erkennbar, dass über «ANGEL WORLD GROUP LIMITED» die kürzeste Verbindung zwischen den beiden Vermittlern verläuft.

Yannic wollte noch herausfinden, welche Vermittler sich alle in der Schweiz niedergelassen haben. Andreas meinte, dass könnte man mit der bestehenden Graph Datenbank einfach ermitteln.

Allerdings braucht es nicht zwingend eine Graph Datenbank um dieses Problem zu lösen. Eine Dokumenten Datenbank könnte genauso gut verwendet werden, da diese ein flexibles Schema bietet (denn bei einigen Vermittlern wurde vergessen der Heimatstaat anzugeben). Somit kann auch effizient nach Vermittlern suchen, welche in der Schweiz ansässig sind.

Tobias hatte die Idee, dass man die Daten auch in eine Time-Series Datenbank laden könnte um nach Schweizer Vermittlern zu suchen.

Doch Andreas fiel ihm sofort ins Wort und meinte, dass er diese Idee gleich Vergessen könne, da Time-Series Datenbanken für Zeitreihen geeignet seien.

Als Beispiel nannte er «Temperaturverlauf» oder «Beschleunigungsmessung».

Da wir allerdings mit Entities und deren Beziehungen arbeiten, kann ein solcher Datensatz praktisch unmöglich in eine Time-Series Datenbank geladen werden.